MCB/BioE/PMB C146/C246
Lecture 18: Finding Motifs II

Scribe: Liana Lareau
Reader: Brant Peterson

March 20, 2003

1 Probabilistic Models

A **probabilistic model** of a sequence has parameters to generate different sequences with different probabilities. A large class of models used to study sequences take the form of **two-component mixture models**. The two components are a set of background frequencies for each letter, assuming that positions in background sequence are independent, and a matrix of frequencies at each position in a motif. The mixing parameter, λ, gives the probability of starting a motif at a position in the sequence. To generate a sequence, at each step either a letter is chosen from the background distribution, or a motif is started and a letter is chosen from each column of the motif model.

For example, a model to generate sequences containing instances of a motif might have the parameters:

background model

A	0.25
C	0.25
G	0.25
T	0.25

motif model

A	1	1	1	1	1	1
C	0	0	0	0	0	0
G	0	0	0	0	0	0
T	0	0	0	0	0	0

mixing parameter $\lambda = 0.01$

This model would generate sequences containing the string `AAAAAA` more often than `CCCCCC`. In theory, such a model generates every possible sequence. One can calculate the likelihood that any given sequence was generated by the model.

Maximum likelihood: With the *data* (sequence) fixed, search the space of *parameters* to find the model which returns our data with the highest probability.

One may also fix the *parameters* and determine the probability that the model generates a given sequence.

2 Philosophy of models

We believe the maximum likelihood parameters tell us something about the biology, *e.g.*, the physical interactions between a binding site and a protein. Another common use for probabilistic models is in gene finding. The goal of such models is not so much to discover the parameters that distinguish coding from non-coding sequence—these do not tell us much about the biology—but to give predictions about new data.

Important caveat: These techniques are models; there is no reason to assume that there is a mapping between *statistical* and *biological* significance.

3 A simple example of motif finding

The full genome sequence and large amount of gene expression data for yeast (*S. cerevisiae*) provided an opportunity to find regulatory motifs for transcription. The yeast genome is about half coding and half non-coding sequence (12 Mb total) with about 6000 genes and only a handful of introns. The promoter regions are small (around 600bp upstream of the gene). They clustered genes based on expression pattern to find groups that were presumably co-regulated, then ran a motif-finding algorithm to find shared sequence in the promoter regions within a group. The first cluster had a very significant, very specific motif that was clearly a transcription factor binding site.

All papers find the same 25-30 motifs in yeast, but the state of the field hasn't progressed very much since then.

Perhaps we are banging our heads on the wrong problem, assuming the model must need more features, and missing something about the biology. Some sites are missed, and there are many statistically significant results that probably have no biological significance (*e.g.*, strings of A's).

We assume there is some biological function relating upstream sequence and gene expression, and a fairly simple model may work in yeast. However, animals are far more complicated. Transcription enhancers in *Drosophila* contain clusters of binding sites for transcription factors that work together, sometimes 10-20 kb from the gene. One attempt to identify enhancers searched for clusters of binding sites, but the occurrence of clusters is simply what one would expect as a result of a Poisson distribution of binding sites; there may be no statistical significance. Nonetheless, the cell is able to use enhancers to control gene expression in a highly specific way.